Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

TC 11 Briefing Papers

NLP-based digital forensic investigation platform for online communications

Dongming Sun[a], *Xiaolu Zhang*[b], *Kim-Kwang Raymond Choo*[b,c], *Liang Hu*[a], *Feng Wang*[a,*]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX, 78249, USA
[c] UniSA STEM, University of South Australia, Adelaide, SA 5095, Australia

ARTICLE INFO

Article history:
Received 23 May 2020
Revised 27 October 2020
Accepted 25 January 2021
Available online 30 January 2021

Keywords:
Digital investigation
Criminal investigation
Email forensics
Social network forensics
NLP-based forensics

ABSTRACT

Digital (forensic) investigations will be increasingly important in both criminal investigations and civil litigations (e.g., corporate espionage, and intellectual property theft) as more of our communications take place over cyberspace (e.g., e-mail and social media platforms). In this paper, we present our proposed Natural Language Processing (NLP)-based digital investigation platform. The platform comprises the data collection and representation phase, the vectorization phase, the feature selection phase, and the classifier generation and evaluation phase. We then demonstrate the potential of our proposed approach using a real-world dataset, whose findings indicate that it outperforms two other competing approaches, namely: LogAnalysis (published in *Expert Systems with Applications*, 2014) and SIIMCO (published in *IEEE Transactions on Information Forensics and Security*, 2016). Specifically, our proposed approach achieves 0.65 in F1-score and 0.83 in precision, whilst LogAnalysis and SIIMCO respectively achieve 0.51 and 0.59 in F1-score and 0.49 and 0.58 in precision.

1. Introduction

Online communication platforms, such as e-mails and social networks, are an important communication and information dissemination platform, and have the potential to be criminally and politically exploited. Recent high profile examples include the relatively recent fake news incidents associated with the U.S. election. For example, Lazer et al. (2018) in their article published in *Science* reported that "fake news stories have gone viral on social media" and emphasized the need for the research community and other stakeholders "to promote interdisciplinary research to reduce the spread of fake news and to address the underlying pathologies it has revealed". Similar observation is reported in the more recent COVID-19 pandemic, as explained in a *Nature* video[1].

There have also been reports and concerns that terrorists and other criminal organizations exploiting social networks to facilitate their illegal activities, such as establishing private

[1] https://www.nature.com/articles/d41586-020-01409-2, last accessed Jul 23, 2020.

* Corresponding author.
E-mail addresses: sundm16@mails.jlu.edu.cn (D. Sun), xiaolu.zhang@utsa.edu (X. Zhang), raymond.choo@fulbrightmail.org (K.-K.R. Choo), hul@jlu.edu.cn (L. Hu), wangfeng12@mails.jlu.edu.cn (F. Wang).
https://doi.org/10.1016/j.cose.2021.102210

communication channels to communicate or disseminate information (Goodman, 2019; Keatinge and Keen, 2019). Hence, it is not surprising that government agencies have also paid close attention to the investigation of such platforms or communication channels. For example in June 2018, the U.S. Department of Homeland Security designed a new online educational training course titled 'Countering Terrorists Exploitation of Social Media and the Internet' to members of the Global Internet Forum to Counter Terrorism (GIFCT)[2].

It can, however, be challenging for (non-authoritarian) governments and service providers to scrutinize and monitor all usages and communications, without infringing on the privacy of the citizens and users. Hence, there have been focus on designing digital investigation techniques for studying different networks, such as social networks (Xu and Chen, 2005). Existing research tends to focus on building a graph of individuals' relationship in a communication network. In such approaches, the key is to identify the most 'closest' associates of a known target (Lee et al., 2018; Liu et al., 2019b). Generally, such approaches seek to achieve improved precision, recall and/or F1-score, but ignore the importance of the content of the conversation/message. Consequently, the approaches can be highly case dependent (designed for specific outcomes), have low accuracy, and are less friendly for digital investigation. For instance, in the shooting rampage at the Gilroy Garlic Festival[3], the shooter had allegedly expressed his anger on his Facebook page prior to the shooting. However, this post did not attract the attention of relevant stakeholders until after the shooting. This is not surprising, as the shooter was not reportedly a known threat on the social network, and hence his post may not be highly weighted using conventional approaches.

Therefore in this paper, we design a digital forensic investigation platform, using Natural Language Processing (NLP), social network vectorization, feature selection, and machine learning (ML) techniques, as the key building blocks. Unlike conventional approaches, the proposed platform focuses on both the relationship between individuals and the content of the communication. Specifically, the platform utilizes the unsupervised NLP model to extract topics from the content of communication messages, and then applies feature selection to rank the topics in order to find the most weighted topics associated with the target individual(s). With the ranked topics, the platform can train classifiers with known generators / algorithms, and the output from the most effective classifier (which can vary between different metrics) is then used to facilitate further investigation.

In the next section, we will review the extant literature. Then in Section 3, we will present the proposed platform, prior to evaluating its performance in Section 4. A comparative summary demonstrates that the proposed platform outperforms two other competing approaches, namely: LogAnalysis by Ferrara et al. (2014) and SIIMCO by Taha and Yoo (2016a). Our proposed platform is also more digital investigation friendly, since using the keywords produced by feature selection can potentially result in additional digital investigation artifacts. We discuss the implications of our research in Section 5. Lastly in Section 6, we conclude this paper.

[2] https://www.dhs.gov/blog/2018/06/11/dhs-announces-launch-countering-terrorists-exploitation-social-media-and-internet, last accessed Jul 23, 2020.

[3] https://www.latimes.com/california/story/2019-07-29/gilroy-garlic-festival-shooting-suspect, last accessed May 23, 2020.

2. Related work

Digital investigation approaches, such as those designed for social network forensics, can be individually-oriented or globally-oriented (Ghani et al., 2018). Individually-oriented digital investigation includes data acquisition and analysis on an individual's computing device such as a mobile device, and mainly focuses on retrieving digital investigation artifacts from the user's account (Arshad et al., 2020; Knox et al., 2020; Shao et al., 2019; Stoyanova et al., 2020). For example, Azfar et al. (2017), Chu et al. (2011), Walnycky et al. (2015), and Norouzizadeh Dezfouli et al. (2016) conducted digital investigation analysis of popular Android and iOS social media applications (apps) and demonstrated the types of artifacts that could be recovered for digital investigation.

Globally-oriented digital investigation, on the other hand, focuses on a broader acquisition and analysis, with the aim of recovering as many implicit relations between individuals as possible (Alqassem et al., 2018; Liu et al., 2019a). For example, Huber et al. (2011) proposed an approach to crawl data from an online social network. Specifically, using Facebook as a case study, the authors demonstrated how data acquisition could be carried out. Globally-oriented analysis could be helpful in identifying influential members of a criminal organization, as well as identifying the leader(s). For example, SIIMCO (Taha and Yoo, 2016a) was designed to build an overview graph that highlights individuals with a strong relationship with known criminals in a social network. Subsequent extensions to this work include optimizing leader identification (Taha and Yoo, 2016b) and communication path identification (Taha and Yoo, 2019). Another related work is LogAnalysis (Ferrara et al., 2014), which was proposed to facilitate criminal detection. LogAnalysis was evaluated using a social network dataset built using phone call records. As our proposed approach is most similar to SIIMCO and LogAnalysis, we will compare the performance of both approaches with our proposed approach (see Section 4.4).

There have also been interest in using artificial intelligence (AI; broadly defined to include both machine and deep learning) techniques to discover suspicious behaviors in a social network. For example, Bindu et al. (2017) proposed an unsupervised learning approach and demonstrated that it is capable of detecting anomalous users automatically from a static social network. However, the assumption is that the structure of the network is not dynamically changed. Hassanpour et al. (2019) used deep convolutional neural networks for images and long short-term memory (LSTM) to extract predictive features of textual data obtained from Instagram. Specifically, they demonstrated that their approach is capable of identifying potential substance use risk behavior, and inform risk assessment and intervention strategy formulation. Tsikerdekis (2016) utilized machine learning to identify deceptive accounts at the time of attempted entry to a online sub-community for prevention. Similarly, the approach

of Ruan et al. (2016) utilizes machine learning to detect compromised accounts based on the online social behaviors of the accounts. Fazil and Abulaish (2018) proposed a hybrid approach for detecting automated spammers in Twitter, using machine learning to analyze the relevant features, such as community-based features (e.g., metadata, content, and interaction-based features). In another independent work, Cresci et al. (2017) utilized machine learning to detect spammers by using digital DNA technology. Specifically, the social fingerprinting technique was designed to discriminate among spambots and genuine accounts in both supervised and unsupervised fashions. Other approaches utilizing AI techniques for different applications include those of Fu et al. (2018) and Shams et al. (2018).

NLP is another technique that has been applied in social network analysis (Al-Zoubi et al., 2017). For example, to verify the owner of a social account, Keretna et al. (2013) utilized a text mining tool (Stanford POS tagger) to extract features from Twitter posts that can represent someone's writing style. The features were then utilized for building a learning module. The approach of Lau et al. (2014) applied both NLP and machine learning techniques on data acquired from Twitter. By testing different NLP and machine learning approaches, Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) reportedly provided the best Area Under the ROC Curve (AUC). Other related works include that of Egele et al. (2017), which was designed to detect compromised accounts on social networks, by analyzing the content of the message and collectively analyzing other features. Anwar and Abulaish (2014) presented a unified social graph based text mining framework to identify digital evidence from chat logs data, based on the analysis of users' conversation and interaction data in the social network. The approach of Wang et al., 2018 considers each HTTP flow generated by mobile apps as a text, utilized natural language processing to extract text-level features and then used the text semantic features of network traffic to develop an effective malware detection model for detecting android malware. The approach presented by Al-Zaidya et al. (2012) was designed to efficiently identify relevant information from a large volume of unstructured textual data, using a systematic method to discover and visualize the criminal networks from documents obtained from a suspects machine. Louis and Engelbrecht (2011) analyzed textual data to discover evidence, utilizing unsupervised information extraction techniques. Such an approach can potentially identify evidence that is missed using a simple keyword search.

3. Our proposed approach

In this section, we will introduce our proposed approach, which comprises the following four phases:

1. **Data collection and representation (see also** Section 3.1): This phase takes as input raw data from a communication network that includes known criminal/noncriminal individuals and their conversations. Therefore, prior to the next phase, the communication network is considered as a set of individuals and their conversations (e.g., a text message or an interactive document), which are labelled as vertices and edges respectively.
2. **Vectorization (see also** Section 3.2): On the premise that the frequency and content of conversations between individuals in a communication network are crucial in determining criminal association(s), in this phase, we utilize an unsupervised NLP model to abstract meaningful topics from those conversations (edges) by which each individual (vertex) is represented with a group of edges that are vectorized with the probability distribution of the topics.
3. **Feature selection (see also** Section 3.3): Since the number of topics involved in conversations can be extremely large, to remove those topics without (significantly) infringing user privacy, we adopt a feature selection algorithm in this phase. This also allows us to enhance the performance of the classifiers generating in the next phase.
4. **Classifier generation and evaluation (see also** Section 3.4): In this final phase, the distribution of the selected topics is used to represent the vertices. Classifiers are generated using known generators/algorithms. According to the metrics of relevance/interest, the most effective classifier can be determined and utilized for future criminal investigation(s).

3.1. *Data collection and representation*

Communication messages posted by individuals involved in the incident (e.g., criminals and their associates) are relevant and differ from those posted by non-criminals. The data collected for this phase is intended to be used to train classifiers. From a practitioner's perspective, the dataset can be built using the data obtained from an existing case and/or prior cases, involving known accounts/abnormal accounts (e.g., human/sex trafficking organizations, organized crime groups terrorists and/or illegal accounts). Sources of the collected conversation data include computing devices used to access the communication network (e.g., mobile devices and apps such as Outlook and Gmail apps), service providers (e.g., Facebook, Twitter and Snapchat), internal communication servers, and an Internet Service Provider (ISP), as long as the data meets the criteria below. Otherwise, data cleaning will be required to ensure that the criteria are met.

- The sender and the receiver involved in a conversation can be identified (e.g., based on the information provided during account registration).
- If a post/message (e.g., a Twitter post and an announcement on a microblog) did not specify a receiver, the sender must be identifiable.
- The conversation captured must be in clear-text or be able to convert to clear-text.
- The criminals/illegal accounts in this communication network must be labeled.

When the raw data is acquired, the proposed approach will represent the communication network with $G = (V, E)$, where V is a set of vertices (or individuals) in the communication network, and E is a set of edges (or conversations) – see Fig. 1.

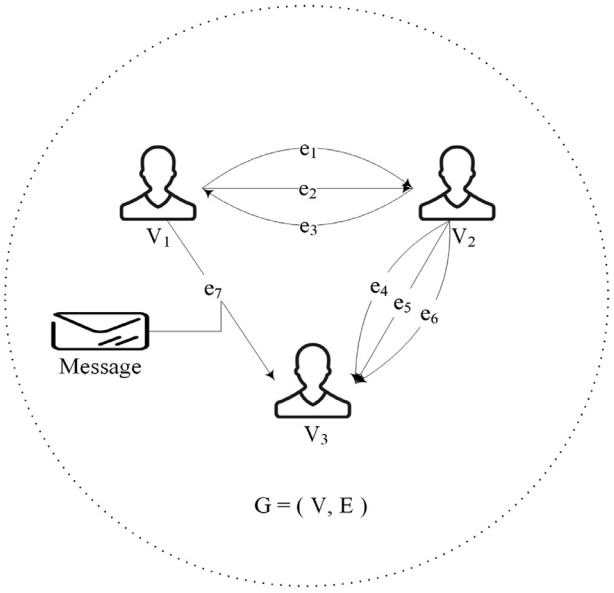


Fig. 1 – Representation of a simple communication network structure in our proposed system.

3.2. *Vectorization*

In this phase, the proposed approach utilizes LDA (Blei et al., 2003) to abstract a number of topics from the conversation (associated with the edges). The distribution of these topics in each edge can be calculated and used for vectorizing the edge. As Eq. 1 shows, when an estimated number of topics t is assigned for building an LDA model, the edge e_i can be converted to vector θ_i that includes the probability (e.g., $P(z_j|e_i)$) of each topic z_j in this edge (or conversation).

$$\theta_i = \left(P(z_1|e_i), P(z_2|e_i), \ldots, P\left(z_j|e_i\right), \ldots, P(z_t|e_i)\right) \quad (1)$$

However, to decide t the perplexity of a LDA model can be used as a reference. This basically indicates how well the model describes the edges. Eq. 2 shows how perplexity is calculated, where $p(w)$ is the probability of a word output from a LDA model and N is the number of the word that appears all over the edges. Therefore, the proposed approach calculates the perplexity for a range of LDA models and then selects one with the lowest perplexity for this vectorization phase.

$$perplexity = e^{-\frac{\sum \log(p(w))}{N}}. \quad (2)$$

Next, as Eq. 3 shows, vertices (e.g., v_k) can be vectorized with the vector of their n edges.

$$v_k = \{\theta_1, \theta_2, \ldots, \theta_n\}. \quad (3)$$

Prior to moving to the next phase, as each vertex may have different number of edges, the proposed approach normalizes each vertex with Eq. 4. The normalized vertex reflects an average distribution of the topics across these edges.

$$v_k = \frac{(\theta_1 + \theta_2 + \ldots + \theta_n)}{n} \quad (4)$$

Table 1 – Vertices before the feature selection.

Samples (vertices)	t_0	t_1	t_2	t_3	t_4	y (label)
v_1	1	0	0	0	0	0
v_2	0	0	0.6	0.1	0.3	1
v_3	0.4	0.2	0.1	0	0.3	0
v_4	0	0.5	0	0	0.5	1

Table 2 – Vertices after the feature selection.

Vectors (vertices)	t_0	t_1	t_3	y (label)
v_1	1	0	0	0
v_2	0	0	0.1	1
v_3	0.4	0.2	0	0
v_4	0	0.5	0	1

3.3. *Feature selection*

As the varying size of the conversations/messages the vectors generated from the last phase can be high-dimensional, the proposed approach utilizes Composition of Feature Relevancy (CFR) (Gao et al., 2018) to reduce the dimensions of v_k. CFR is a feature selection algorithm based on Mutual Information (MI[4]) that can discover the impact for each topic (hereafter, a topic is also considered as a feature for 'feature selection'). Eq. 5 (Gao et al., 2018) shows how CFR works. For readers who are unfamiliar with CFR, the sample data is given in Tables 1 and 2.

Table 1 includes a group of sample vertices (from v_1 to v_4) that are represented by topics from t_0 to t_4. Each topic has a distribution for these vertices. Assuming that the desired number of selected features is set to three, the features' distribution in Table 1 will be sent to CFR (Eq. 5) as the input X_k and, then the features which are ranked having the largest output $J(X_k)$ will be moved to the selected feature set. This process will be repeated twice on the remaining features until all three features are found.

For instance, as the selected feature subset S started empty Eq. 5 and was identical to $J(X_k) = I(X_k; y)$ until the first feature was found from Table 1. As label Y (which indicates whether the vertex is criminal – 1 or non-criminal – 0) was known, $J(X_k)$ can be calculated for each feature as $J(t_0) = J(t_1) = 1.0 > J(t_2) = J(t_4) = 0.67 > J(t_3) = 0.19$. Therefore, t_0 was the first feature added to the feature subset. Similarly, t_3 was selected as the second feature since $J(t_3) = -0.19 > J(t_2) = J(t_4) = -0.67 > J(t_1) = -1.0$ and t_1 was selected as the third feature since $J(t_1) = J(t_4) = 0.38 > J(t_2) = -0.62$. Thus, the selected feature set was created and shown in Table 2.

$$J(X_k) = \sum_{X_j \in S} \left\{ I(X_k; Y|X_j) - I(X_k; Y; X_j) \right\} \quad (5)$$

Since the size of the selected feature set may impact the performance of the classifier, our proposed approach builds

[4] In information theory, MI is utilized for measuring the interdependence among variables.

classifiers with every possible number of features and then sorts out the optimal classifier that can be utilized – see Section 3.4.

3.4. *Classifier generation & evaluation*

With the selected feature set obtained from the preceding phase, the proposed approach now generates a group of classifiers (again, the proposed approach is designed to find the most effective classifier for the given communication network) with commonly used classifier generators, such as AdaBoost (AB), Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), Random Forest(RF) and Support Vector Machines(SVM).

In the proposed approach, we use 'scikit-learn', an existing Machine-Learning module in Python, in which those classifier generators needed were pre-implemented. The process for training the classifiers is typical and programmatically standardized. An example for creating an AdaBoost classifier with 'scikit-learn' is given in the following steps:

1. Initialize the generator:

```
from sklearn.ensemble import
    AdaBoostClassifier
clf_ab = AdaBoostClassifier()
```

2. Train the classifier *clf_ab* with a selected feature set (e.g., data in Table 2):

```
clf_ab.fit(v, y)
```

In the above equation, v and y denote the vectors (e.g., v_1-v_3 in Table 2) and their label (e.g., the value in the column y), respectively. Function $fit(.)$ trains a classifier with the data input.

Again, to find the most effective classifier, the proposed approach evaluates them in terms of metrics such as precision, recall and F1-score. However, by default, F1-score is considered our primary metric as it is a more comprehensive measure that considers both the precision and the recall. Since the calculation for these values is well-known, we omitted the discussion of these metrics.

In addition, to apply a classifier to facilitate criminal investigation, the test data must go through the vectorization process first (see Section 3.2). Then, the vector of each vertex can be loaded to function *predict*(..) of the instance of this classifier. As the listing below shows, label *y_pre* (criminal/noncriminal) can be predicted for a vertex *x_test*.

```
y_pre=clf_ab.predict(x_test)
```

4. Evaluation and findings

In this section, we explain our evaluation setup and the findings. First, we validate the performance of the proposed approach using a real-world dataset (see Sections 4.1 and 4.2). Then, we compare the performance of the proposed approach with and without feature selection using the same dataset. The findings highlight the necessity of feature selection (see Section 4.3). Lastly, we compare the proposed approach with LogAnalysis (Ferrara et al., 2014) and SIIMCO (Taha and Yoo, 2016a), and present the comparative summary of their performance and that of our proposed approach (see Section 4.4).

4.1. *Dataset & environment setup*

We used the real-world dataset from the Enron Email Corpus (Enron Email Dataset, 0000; Keila and Skillicorn, 2005). To clean the dataset, we preprocessed the dataset by removing duplicates, junk mails, undelivered and empty Emails and punctuation marks (for applying LDA). Thus, in total, 47,468 emails sent/received from/to 166 former Enron employees remained, in which 25/166 employees were confirmed 'criminals' (individuals allegedly found to be involved in the fraudulent activities). For reproducibility, we post both the original dataset and the processed dataset on our GitHub repository[5].

Our proposed approach applies a transductive method, that is the LDA model is applied on the entire dataset. Then, as the dataset was considered relatively imbalanced (only 25/166 employees are criminals), the framework was evaluated across a 5∗2 Nested Cross Validation, in which the preprocessed dataset was split into five folds. Each fold can potentially be chosen to be the test set and the remaining four were then used for another 2-fold validation, where if 1 fold became a training set the other must become the validation set. The training set was used for classifier training, in which each generator built a group of classifiers for each possible number of topics which was larger than 0 and less than or equal to the number given by the LDA who gained the smallest perplexity. The validation set was used for testing these classifiers in terms of precision, recall and F1-score. Only the best classifiers for each metrics were recommended to the investigator and evaluated in the test set.

The evaluation was performed on a Ubuntu 5.4.0 PC with 16GB RAM, and our proposed approach was implemented using Python.

4.2. *Performance validation*

As Fig. 2 shows, we found that the LDA model gained the smallest perplexity and was built with 210 topics (due to our prior experience, 10 to 250 topics were tested). Thus, prior to classifier generation, the given dataset was vectorized with this 210-topic LDA model.

Next, we went through the 5∗2 Nested Cross Validation. For each of the outer 5 folds, each classifier generator built classifiers for all possible top n ($1 \leq n \leq 210$) topics (that were ranked by CFR) from the training set (thus, 210 classifiers were created per generator). Then, the 210 classifiers generated were applied to the validation set in order to find those with the best performance. We remarked that in a real-world scenario,

[5] https://github.com/Sun121sun/ENRON-EMAILS-AND-EMPLOYEE.

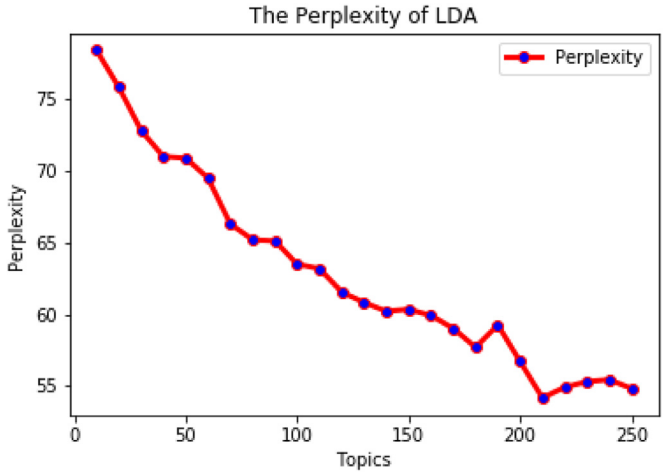


Fig. 2 – The perplexity of LDA models.

Table 3 – The performance for each type of classifier in each fold.

Fold	Metrics	DT	NB	AB	LR	RF	SVM
Fold 1	Precision	(0.5484,129)	(0.7143,200)	(0.6154,78)	(**0.9375**,67)	(0.8735,88)	(0.5376,86)
	Recall	(0.6400,118)	(**0.9600**,75)	(0.6000,60)	(0.6000,52)	(0.5800,78)	(0.6237,93)
	F1-score	(0.5556,177)	(0.5063,72)	(0.5714,95)	(**0.7500**,182)	(0.6700,162)	(0.6683,72)
Fold 2	Precision	(0.6087,102)	(0.2432,57)	(0.4118,46)	(0.5625,105)	(**0.7936**,83)	(0.3944,80)
	Recall	(0.4400,132)	(**0.6400**,75)	(0.4000,135)	(0.3600,103)	(0.4937,77)	(0.5453,102)
	F1-score	(0.5833,181)	(0.4545,70)	(0.4167,111)	(0.4500,135)	(**0.6071**,35)	(0.4725,83)
Fold 3	Precision	(0.5000,172)	(0.5882,180)	(0.5588,108)	(0.7692,52)	(**0.8920**,63)	(0.5929,99)
	Recall	(0.5200,120)	(**0.9200**,35)	(0.6800,119)	(0.6400,112)	(0.7223,97)	(0.7369,105)
	F1-score	(0.5714,151)	(0.5000,191)	(**0.7556**,94)	(0.6511,153)	(0.6928,128)	(0.6638,175)
Fold 4	Precision	(0.4857,145)	(0.6875,206)	(0.5000,112)	(0.7500,89)	(**0.8333**,107)	(0.5121,79)
	Recall	(**0.7600**,179)	(**0.7600**,81)	(0.6400,157)	(0.6800,104)	(0.6635,93)	(0.6398,124)
	F1-score	(0.5818,117)	(0.5366,195)	(0.5714,138)	(0.6883,136)	(**0.7226**,65)	(0.5939,19)
Fold 5	Precision	(0.5263,183)	(0.3947,54)	(0.5600,83)	(0.6818,96)	(**0.8367**,78)	(0.4398,81)
	Recall	(0.4000,124)	(**0.7600**,78)	(0.6000,94)	(0.6000,95)	(0.5853,100)	(0.6599,128)
	F1-score	(0.4889,142)	(0.5455,194)	(0.5490,158)	(0.6522,90)	(**0.6971**,158)	(0.5369,90)

Table 4 – The average precision, recall and F1-score for each type of classifier through the Nested Cross Validation.

Metrics	DT	NB	AB	LR	RF	SVM
Average Precision	0.5338	0.5256	0.5292	0.7402	**0.8332**	0.5169
Average Recall	0.5520	**0.8080**	0.5840	0.5760	0.5932	0.6321
Average F1-score	0.5562	0.5086	0.5728	0.6223	**0.6500**	0.5634

selected classifiers should be classifiers that are applied to an unlabeled dataset for criminal prediction. As Table 3 shows, we listed 3 classifiers per generator per fold, which had the best F1-score, recall and precision. For each classifier, we also included the number of the ranked features utilized for their generation.

In Table 4, we further calculated the average precision, recall and F1-score for these generators, in which NB gained the best average recall (0.81), and RF had the best average precision (0.83) and the best average F1-score (0.65).

4.3. *Feature selection vs. no-feature selection*

We posited that feature selection plays a significant role in our proposed approach outperforming the two other competing approaches and conventional approaches, we intended to test how the proposed approach works without feature selection. To exclude feature selection, the classifier generators must use the entire feature set (210 features in this case) obtained from LDA. Thus, we simply utilized 5-fold evaluation to acquire the average precision, recall and F1-score for these classifiers and compared them with the classifiers built with feature selection in Table 5.

The comparison shows that except for the recall of AB which had the same score, the proposed approach with feature selection outperforms for every metrics in each classifier. If we compare the highest score, feature selection helped NB achieves a 0.38 higher recall than NB on the non-feature selec-

Table 5 – Feature selection vs. none-feature selection.

	Metrics	DT	NB	AB	LR	RF	SVM
w/ Feature selection	Precision	0.5338	0.5256	0.5292	0.7402	**0.8332**	0.5169
	Recall	0.5520	**0.8080**	0.5840	0.5760	0.5932	0.6321
	F1-score	0.5562	0.5086	0.5728	0.6423	**0.6500**	0.5634
w/o feature selection	Precision	0.5192	0.5065	0.5152	0.7011	**0.7210**	0.3463
	Recall	0.5060	0.4240	**0.5840**	0.5680	0.4356	0.5720
	F1-score	0.5261	0.4669	0.5471	0.5963	**0.5997**	0.4534

Table 6 – Performance of SIIMCO and LogAnalysis: A comparative summary.

Approach	F1-score	Precision	Recall
LogAnalysis[a]	0.51	0.49	0.53
SIIMCO[a]	0.59	0.58	0.60
Our proposed approach	0.65 (RF)	0.83 (RF)	0.59 (RF)

[a] The results are obtained from (Taha and Yoo, 2016a, Fig.4). All the three approaches are evaluated on a common Enron Email dataset.

tion. Similarly, feature selection achieved 0.11 higher precision and 0.05 higher F1-score.

4.4. *Proposed approach vs. existing works*

In this section, we compared the performance of our proposed approach with SIIMCO (Taha and Yoo, 2016a) and LogAnalysis (Ferrara et al., 2014). For consistency, we evaluated the proposed approach using a common Enron Email dataset[6], since SIIMCO and LogAnalysis were reportedly evaluated by (Taha and Yoo, 2016a, Fig. 4) using the same dataset. From Table 6, we observe that the proposed approach outperforms both SIIMCO and LogAnalysis. For example, when all three approaches were evaluated using the same dataset, our proposed approach achieves a higher F1-score and precision rate at 0.06 and 0.25,respectively.

5. Discussion

In this section, we discuss the findings detailed in the preceding section. Specifically, we focus on the classifier selection, the necessity of feature selection and the hidden knowledge associated with our proposed approach and the existing approaches.

5.1. *Classifier selection*

A classifier may perform differently among metrics. As Table 4 shows, RF had the best average F1-score as well as three out of the five folds during the cross-validation (see Table 3). In the three winning folds, a 0.02 to 0.21 higher F1-score was observed.

[6] The raw dataset can be found from http://www.cs.cmu.edu/~enron/ and https://github.com/Sun121sun/ENRON-EMAILS-AND-EMPLOYEE.

Although we recommend using F1-score as the default metrics, our proposed approach can be customized by the digital investigators, based on their preference and case-specific needs. For example, as Table 3 shows, RF would be a good choice as RF had the best precision in 4 out of 5 folds (roughly 0.08to 0.18 higher than the second high classifier). On the other hand, however, if the goal for an investigation was to find as many suspects as possible, then recall should be the metric of interest. Therefore, based on our findings, NB would be a good candidate since it had the best average in every fold (0to 0.32 higher than the second high classifier).

5.2. *Benefits of feature selection*

Our proposed approach benefits from feature selection in two aspects. First, as Table 5 shows, feature selection results in significant improvement for Precision, Recall and F1-score. Second, using perplexity to determine that the number of features(topics) is still large. From a digital investigator's perspective, feature selection can help facilitate the sorting of topics (not) relevant to the case. We argue that such sorted topics/keywords may also facilitate in the discovery of additional digital investigation artifacts. For example, Table 7 shows the top 15 features (topics) obtained from one of the five folds in the Nested Cross Validation. By comparing the employees' name list with the known criminal list, we determined that 19 (of the 40 names included in these top sorted topics) are those of the criminals (first/last). In addition to criminal names, feature selection could potentially retrieve additional digital investigation artifacts such as location, address or timestamp.

Another benefit of applying sorted topics in digital investigations is that the sorted topic can be used as a triage tool, in the sense that red flag indicators may be identified from the significant volume of data in the communication networks. As an example, we ranked the Emails in the test set of a fold by the probability of topic 106 and then manually analyzed the content of the top 50 Emails – see Fig. 3. Unsurprisingly, one of the emails (Email Message-ID: 19471882.1075854486799.JavaMail.evans@thyme) shown in Fig. 5 was flagged. Specifically, the email content is about a conversation between 'suspect 59' (anonymized due to privacy concerns; a former super executive of Enron Corporation), and 'suspect 28' (a former super executive of Enron North America and Enron Energy Services), who were reportedly sentenced to 24 years' imprison-

Table 7 – The selected topics.

Topic	words
177	Oct. sells Sept. Carroll 'suspect 79' 'suspect 97' Ronald 'suspect 79' 'suspect 118' 'suspect 118' derrick 'suspect 59'
106	exploration evergreen cancellation CFO solicit amending improve receivable recommendations advertisements February polling
27	enronxgate 2001 Williams original message 2000 sent 'suspect 97' 'suspect 66' subject Shively 'suspect 59'
75	exhibit pinch restricted ability bankers enforceability deviate emphasizing tenor suggestions points penalized
5	arbitration checklist 'suspect 6' traveling pushes official pig Phibro frequent relate shareholder fiduciary
133	Vancouver disputes behavior conduct compensation court Pennsylvania zones operates emergency injunction supplemental
41	2000 Jones Tanya Brant forwarded 10pm credit Europe global Clark Lisa 'suspect 28'
181	worksheet Julie Jacoby hotmail Johnson 'suspect 6' Fred Shively yahoo Kelly Baughman enron discuss
112	threshold dated agreement amendment sheet July west executed 2001 schedule Canadian distributed
3	'suspect 80' 'suspect 81' Monday tomorrow original sent thanks know call leave meeting Friday
63	directory plastics generic joy gold Cara gwhalle queue biological publications audience wondering
135	gift 'suspect 2' ancillary Marvin incorporation breakout indemnification certificate 'suspect 6' 'suspect 12' directors 'suspect 118'
204	request rock accomplishments goals scores link original units message allocations resource piper
66	phase 'suspect 66' integration incorporated Sunday enrononline exchanges 7th kitchen Saturday plans presto
149	wrong wrestling ownership bill smurfit Thanx Sundance discrepancy real alternative time scheduling

Words in Red refer to the (first/last) name of criminals (replaced with code-name for privacy concerns).
Words in Blue refer to a location.
Words in Green refer to a year/month/date.
(*Although the dataset is a publicly available, we remove the names of the individuals in this paper to avoid potential privacy implications.)

Words in Red refer to the (first/last) name of criminals (replaced with code-name for privacy concerns).
Words in Blue refer to a location.
Words in Green refer to a year/month/date.
(*Although the dataset is a publicly available, we remove the names of the individuals in this paper to avoid potential privacy implications.)

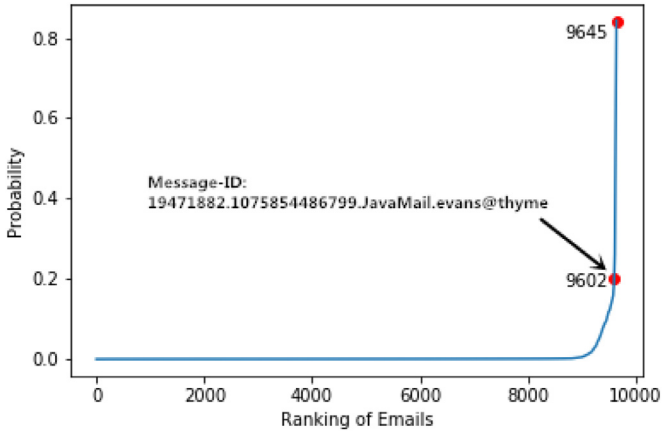


Fig. 3 – The full dataset ranked by the probability of topic 106.

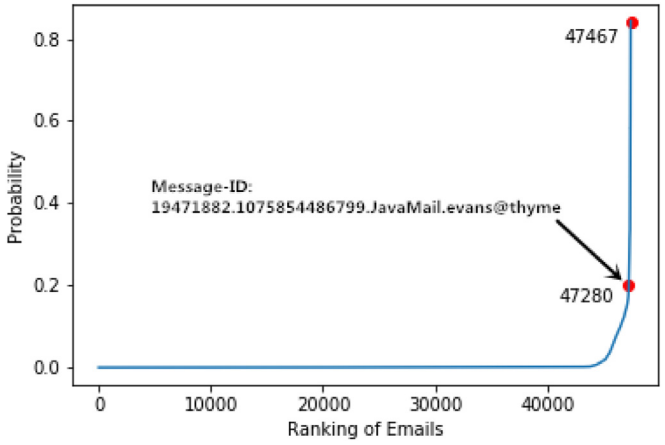


Fig. 4 – The test dataset of one fold ranked by the probability of topic 106.

Message-ID: <19471882.1075854486799.JavaMail.evans@thyme>

Date: 2000/8/28 12:47

From: Suspect 28@enron.com

To: Suspect 59@enron.com

Subject: Cangen

X-From: 'Suspect 28'

X-To: 'Suspect 59'

'Suspect 59',

you will be receiving a DASH from the principal investing group which does not fit well into some of the charter criteria; however, I do think this is a venture we should support. In a nutshell, it is providing $10MM, in seed capital, to be used in the start up of **a new distributed generation company having the primary mission of designing, fabricating and marketing 2MW to 22MW mobile turbine-generator sets**. These units are fully mobile via the nations highways and can meet single digit NOX requirements given current SCR technology. They are primarily utilized for back-up power, emergency power, VAR creation and peaker application and are fully dual fuel capable. The ultimate vision is a fleet of mobile/flexible/low cost generator sets that can be put in place very quickly in the most stringent emissions and constrained electrical regions and relocated when market conditions change. The benefits include: a) 50% plus interest in a new distributed generation company with the opportunity to raise private funds at later rounds and ultimately the possibility of an IPO - similar companies in this space include Active Power, Capstone and Elektryon; b) origination opportunities - the ENA mid-stream origination groups believe that there are numerous opportunities (2001 to 2002 time frame) to utilize this technology to take advantage of market opportunities and customer applications; c) learning - the distributed generation/renewable space is an area of critical importance for ENA and EES to understand. If ENA through this venture can make this technology work, it could be a very powerful tool in our network used to capture opportunity and manage risk; d) low risk - it is expected that roughly 50% of the initial seed capital would be recoverable through asset sales if for some reason the technology did not work or the market did not materialize; e) low tech - this solution utilizes existing turbine/SCR technology; and f) ENA would have a controlling position in this new venture. However, this venture is different than Active Power and several other Principal Investing investments in that a) it is a very early stage company and b) ENA will have to provide or procure a complete management team if the technology proves out. Given, that we could benefit most from the successful application of this venture, it is worth the initial work. Our partner is Power Systems Manufacturing of Boca Raton, Florida. We have a very solid history with this company. They have successfully solved many of the technical issues surrounding the Korean 6B's which are currently operating at New Albany. They would bring the technical and engineering expertise at cost for their 50%. ENA would ultimately have a buy-out clause at a fixed number for their 50%. If you would like to discuss in greater detail do not hesitate to call and I will set up a meeting.

Regards

'Suspect 28'

Fig. 5 – The evidential Email whose important was indicated by topic 106.

ment and 2.5 years' imprisonment, respectively[7,8]. In addition, it was also reported that 'suspect 28' "turned over his $4.2 million in illegal trading profits to the Justice Department, and another $3 million to the Securities and Exchange Commission". Contents from the Email we located, as shown in Fig. 5, also showed that 'suspect 28' was advocating the use of "ENA", a newly established company, to 'suspect 59'. In a number of U.S. states, business owners are required to obtain a business license. Therefore, it is easy for the recipient, in this case, 'suspect 59' to search and find out the beneficial owner of this business who happened to be 'suspect 28'. Hence, together with other information / leads, this email could be of digital investigative interest since one would know when the conversation about this matter commenced. In addition, to show if this topic is applicable for a larger scale of data we applied the same approach to the entire dataset, comprising 47,468 Emails. As shown in Fig. 4, as the scale of data increases, the same email (i.e., Fig. 5) was also located within the top 200 emails.

5.3. *A comparative summary*

Our proposed approach is the first work to integrate LDA, feature selection and machine learning to online communication digital investigation (and in this paper, we used email communication as the use case).

One reason why there were significant true/false negatives in LogAnalysis and SIIMCO is because both approaches were built based on creating a overall graph of suspects' relationship, where a highly active user in the communication network was usually over-weighted as a criminal. As an example, 'employee 57', during our investigation, who had sent and received 3247 and 847 Emails respectively was considered a criminal in both existing approaches (i.e., a true negative). Our proposed approach, on the other hand, focuses not only on the relationship but also the communication content. In other words, a user would be labeled as a criminal only if the message (s)he sent/received were remarkably similar to a message involved in a known criminal. Thus, based on ML techniques, our proposed approach could provide a more reliable finding. Using the same example, 'employee 57', was determined not to be a criminal in our proposed approach.

Our proposed approach consists of features that can benefit digital investigators. For example, if the first priority of a case is to cast a wide net and identify as many as suspects as possible, the investigator should consider using the classifier with the best recall. In the event that the aim is to obtain the most reliable list of suspects, the investigator should consider choosing the classifier with the best precision. And again, as discussed in Section 5.2, feature selection can benefit digital investigation by providing these highly suspicious keywords.

6. Conclusion and future work

As communications become digitalized, online communication channels such as emails will become an increasingly im-

[7] http://www.washingtonpost.com/wp-dyn/content/article/2009/10/13/AR2009101300782.html, last accessed Jul 13, 2020.

[8] https://www.chron.com/business/enron/article/Judge-decries-Enron-exec-s-deeper-guilt-1889810.php, last accessed Jul 13, 2020.

portance source of evidence in criminal investigations (e.g., fake news and foreign influence in an election) and civil litigations (e.g., e-discovery and defamation).

In this paper, we presented a ML-based platform for online communication digital investigation, which also integrates feature selection and NLP. Using a real-world dataset, we demonstrated its utility as well as how it outperforms two other competing approaches, namely: LogAnalysis (Ferrara et al., 2014) and SIIMCO (Taha and Yoo, 2016a).

Future research will include evaluating our proposed approach using other datasets, such as the recent COVID-19 related cyber criminal activities (e.g., frauds). This will also allow us to demonstrate the generalizability of our proposed approach, for example how the approach can be used to identify person(s)-of-interest.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Dongming Sun: Methodology, Formal analysis, Writing - original draft. **Xiaolu Zhang:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Kim-Kwang Raymond Choo:** Conceptualization, Methodology, Writing - review & editing. **Liang Hu:** Conceptualization, Methodology, Project administration, Resources, Supervision, Writing - review & editing, Funding acquisition. **Feng Wang:** Conceptualization, Methodology, Project administration, Resources, Supervision, Writing - review & editing, Funding acquisition.

Acknowlgedgments

The authors thank the handling editor and the three reviewers for their constructive feedback. Despite their invaluable assistance, any errors remaining in this paper are solely attributed to the authors. This work is also funded by: National Key R&D Plan of China under Grant No. 2017YFA0604500, and by National Sci-Tech Support Plan of China under Grant No. 2014BAH02F00, and by National Natural Science Foundation of China under Grant No. 61701190, and by Youth Science Foundation of Jilin Province of China under Grant No. 20180520021JH, and by Key Technology Innovation Cooperation Project of Government and University for the whole Industry Demonstration under Grant No. SXGJSF2017-4, and by Key scientific and technological R&D Plan of Jilin Province of China under Grant No. 20180201103GX, and by Project of Jilin Province Development and Reform Commission under Grant No. 2019FGWTZC001. K.-K. R. Choo was supported only by National Science Foundation CREST under Grant HRD-1736209, and the Cloud Technology Endowed Professorship.

REFERENCES

Al-Zaidya R, CMFung B, Youssef AM, Fortin F. Mining criminal networks from unstructured text documents. Digital Invest. 2012;8(3–4):147–60.

Al-Zoubi A, Paris H, Alqatawna J. Spam profile detection in social networks based on public features. In: 2017 8th International Conference on information and Communication Systems (ICICS). Irbid Jordan: IEEE; 2017. p. 130–5.

Alqassem I, Rahwan I, Svetinovic D. The anti-social system properties: Bitcoin network data analysis. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2018:1–11.

Anwar T, Abulaish M. A social graph based text mining framework for chat log investigation. Digital Invest. 2014;11(4):349–62.

Arshad H, Jantan A, Hoon GK, Abiodun OI. Formal knowledge model for online social network forensics. Comput. Secur. 2020;89.

Azfar A, Choo K-KR, Liu L. Forensic taxonomy of android social apps. J. Forensic Sci. 2017;62(2):435–56.

Bindu P, Thilagam PS, Ahuja D. Discovering suspicious behavior in multilayer social networks. Comput Human Behav 2017;73:568–82.

Blei DM, Ng AY, Jordan MI, Lafferty J. Latent dirichlet allocation. Journal of Machine Learning Research 2003;3:993–1022.

Chu H-C, Deng D-J, Park JH. Live data mining concerning social networking forensics based on a facebook session through aggregation of social data. IEEE J. Sel. Areas Commun. 2011;29(7):1368–76.

Cresci S, Pietro RD, Petrocchi M, Spognardi A, Tesconi M. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. IEEE Transactions on Dependable & Secure Computing 2017;15(4). doi:10.1109/TDSC.2017.2681672.

Egele M, Stringhini G, Kruegel C, Vigna G. Towards detecting compromised accounts on social networks. IEEE Transactions on Dependable & Secure Computing 2017;14(4):447–60.

Enron Email Dataset, Enron Email Dataset. [Online].

Fazil M, Abulaish M. A hybrid approach for detecting automated spammers in twitter. IEEE Trans. Inf. Forensics Secur. 2018;13(11):2707–19.

Ferrara E, Meo PD, Catanese S, Fiumara G. Detecting criminal organizations in mobile phone networks. Expert Syst. Appl. 2014;41(13):5733–50.

Fu K, Chen Z, Lu C-T. StreetNet: preference learning with convolutional neural network on urban crime perception. In: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle Washington: ACM; 2018. p. 269–78.

Gao W, He J, Hu L, Zhang P. Feature selection considering the composition of feature relevancy. Pattern Recognit Lett 2018;112:70–4.

Ghani NA, Hamid S, Hashem IAT, Ahmed E. Social media big data analytics: a survey. Comput Human Behav 2018;101:417–28.

Goodman AEJ. When you give a terrorist a twitter: holding social media companies liable for their support of terrorism. Pepperdine Law Rev 2019;46(1):147–202.

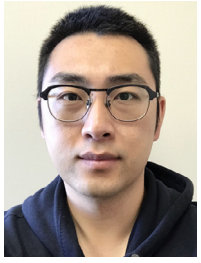
Hassanpour S, Tomita N, DeLise T, Crosier B, Marsch LA. Identifying substance use risk based on deep neural networks and instagram social media data. Neuropsychopharmacology 2019;44(3):487–94.

Huber M, Mulazzani M, Leithner M, Schrittwieser S, Wondracek G, Weippl E. Social snapshots: Digital forensics for online social networks. In: Proceedings of the 27th Annual Computer Security Applications Conference. Orlando Florida USA: ACM; 2011. p. 113–22.

Keatinge T, Keen F. Social media and (counter) terrorist finance:

afund-raising and disruption tool. Studies in Conflict & Terrorism 2019;42(1–2):178–205.
Keila PS, Skillicorn DB. Structure in the enron email dataset. Comput. Math. Org. Theory 2005;11(3):183–99.
Keretna S, Hossny A, Creighton D. Recognising user identity in twitter social networks via text mining. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics. Manchester UK: IEEE; 2013. p. 3079–82.
Knox S, Moghadam S, Patrick K, Phan A, Choo K-KR. Whats really happning? a forensic analysis of android and iOS happn dating apps. Computers & Security 2020:101833.
Lau RY, Xia Y, Ye Y. A probabilistic generative model for mining cybercriminal networks from online social media. IEEE Comput Intell Mag 2014;9(1):31–43.
Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, et al. The science of fake news. Science 2018;359(6380):1094–6.
Lee E, Woo J, Kim H, Kim HK. No silk road for online gamers!: using social network analysis to unveil black markets in online games. In: Proceedings of the 2018 World Wide Web Conference. Lyon France: International World Wide Web Conferences Steering Committee; 2018. p. 1825–34.
Liu B, Zhou Q, Ding R-X, Palomares I, Herrera F. Large-scale group decision making model based on social network analysis: trust relationship-based conflict detection and elimination. Eur J Oper Res 2019;275(2):737–54.
Liu W, Gong D, Tan M, Shi Q, Yang Y, Hauptmann AG. Learning distilled graph for large-scale social network data clustering. IEEE Trans Knowl Data Eng 2019.
Louis AL, Engelbrecht AP. Unsupervised discovery of relations for analysis of textual data.. Digital Invest. 2011;7(3):154–71.
Norouzizadeh Dezfouli F, Dehghantanha A, Eterovic-Soric B, Choo K-KR. Investigating social networking applications on smartphones detecting facebook, twitter, linkedin and Google+ artefacts on android and iOS platforms. Aust. J. Forensic Sci. 2016;48(4):469–88.
Ruan X, Wu Z, Wang H, Jajodia S. Profiling online social behaviors for compromised account detection. IEEE Trans. Inf. Forensics Secur. 2016;11(1):176–87.
Shams S, Goswami S, Lee K, Yang S, Park S-J. Towards distributed cyberinfrastructure for smart cities using big data and deep learning technologies. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). Vienna Austria: IEEE; 2018. p. 1276–83.
Shao S, Tunc C, Al-Shawi A, Hariri S. Automated twitter author clustering with unsupervised learning for social media forensics. In: 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). IEEE; 2019. p. 1–8.
Stoyanova M, Nikoloudakis Y, Panagiotakis S, Pallis E, Markakis EK. A survey on the internet of things (iot) forensics: challenges, approaches and open issues. IEEE Communications Surveys & Tutorials 2020.
Taha K, Yoo PD. SIIMCO: A forensic investigation tool for identifying the influential members of a criminal organization. IEEE Trans. Inf. Forensics Secur. 2016;11(4):811–22.
Taha K, Yoo PD. Using the spanning tree of a criminal network for identifying its leaders. IEEE Trans. Inf. Forensics Secur. 2016;12(2):445–53.
Taha K, Yoo PD. Shortlisting the influential members of criminal organizations and identifying their important communication channels. IEEE Trans. Inf. Forensics Secur. 2019;14(8):1988–99.
Tsikerdekis M. Identity deception prevention using common contribution network data. IEEE Trans. Inf. Forensics Secur. 2016;12(1):188–99.
Walnycky D, Baggili I, Marrington A, Moore J, Breitinger F. Network and device forensic analysis of android social-messaging applications. Digital Invest. 2015;14:S77–84.
Wang S, Yan Q, Chen Z, Bo Y, Zhao C, Conti M. Detecting android malware leveraging text semantics of network flows. IEEE Trans. Inf. Forensics Secur. 2018;13(5):1096–109.
Xu JJ, Chen H. Crimenet explorer: a framework for criminal network knowledge discovery. ACM Trans. Inf. Syst. 2005;23(2):201–26.

Dongming Sun received his M.S. degree in the College of Software from Jilin University in 2013. He is working toward the Ph.D. degree in the College of Computer Science, Jilin University. His research interests include social network forensic analysis.



Xiaolu Zhang received the Ph.D. degree in Computer Science from Jilin University, Changchun, China, in 2016. He was a Visiting Ph.D. Student with the University of New Haven, West Haven, CT, USA. He is currently an Associate Professor with the University of Texas at San Antonio (UTSA). His current research interests include cyber security and digital forensics. He was also the recipient of a China Scholarship Council Scholarship for his doctoral work and UTSA's Endowed 1969 Commemorative Award for Teaching Excellence.

Kim-Kwang Raymond Choo holds the Cloud Technology Endowed Professorship at The University of Texas at San Antonio (UTSA), and has a courtesy appointment at the University of South Australia.. In 2016, he was named the Cybersecurity Educator of the Year - APAC, and in 2015 he and his team won the Digital Forensics Research Challenge organized by Germany's University of Erlangen-Nuremberg. He is an IEEE Computer Society Distinguished Visitor (2021 - 2023), and included in Web of Science's Highly Cited Researcher in the field of Cross-Field - 2020. He is also the recipient of the 2019 IEEE TC on Scalable Computing Award for Excellence in Scalable Computing (Middle Career Researcher), the 2018 UTSA College of Business Col. Jean Piccione and Lt. Col. Philip Piccione Endowed Research Award for Tenured Faculty, the British Computer Society's 2019 Wilkes Award Runner-up, the 2014 Highly Commended Award by the Australia New Zealand Policing Advisory Agency, the Fulbright Scholarship in 2009, the 2008 Australia Day Achievement Medallion, and the British Computer Society's Wilkes Award in 2008. He has also received best paper awards from the IEEE Consumer Electronics Magazine for 2020, EURASIP JWCN in 2019, IEEE TrustCom 2018, and ESORICS 2015; the Korea Information Processing Society's JIPS Most Cited Paper Award for 2020 and Survey Paper Award (Gold) in 2019; the IEEE Blockchain 2019 Outstanding Paper Award; and Best Student Paper Awards from Inscrypt 2019 and ACISP 2005.

Liang Hu received his M.S. and Ph.D. degrees in Computer Science from Jilin University in 1993 and 1999. He is currently a full Professor and doctoral supervisor at the College of Computer Science and Technology, Jilin University, China. His research areas are network security and distributed computing, including the theories, models, and algorithms of PKI/IBE, IDS/IPS, and grid computing. He is a member of the China Computer Federation.

Feng Wang received his M.S. and Ph.D. degrees in Computer Science from Jilin University in 2012 and 2016 respectively. He is currently an Associate Professor in Jilin University. His research interests include computer networks, information security, the Internet of Things and Cyber Physical Systems.